Computers in Biology and Medicine 113 (2019) 103397

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Effective hybrid approach for protein structure prediction in a two-dimensional Hydrophobic–Polar model

Cheng-Hong Yang[a,b], Yu-Shiun Lin[a], Li-Yeh Chuang[c,d,*], Yu-Da Lin[a,**]

[a] Department of Electronic Engineering, National Kaohsiung University of Science and Technology, No.1, Sec. 1, Syuecheng Rd., Dashu District, Kaohsiung City, 84001, Taiwan
[b] Program in Biomedical Engineering, Kaohsiung Medical University, No.100, Tzyou 1st Rd., Sanmin Dist., Kaohsiung City, 80756, Taiwan
[c] Department of Chemical Engineering, I-Shou University, No.415, Jiangong Rd., Sanmin Dist., Kaohsiung City, 807, Taiwan
[d] Institute of Biotechnology and Chemical Engineering, I-Shou University, No.415, Jiangong Rd., Sanmin Dist., Kaohsiung City, 807, Taiwan

ARTICLE INFO

Keywords:
Two-dimensional hydrophobic–polar model
Protein folding prediction
Protein sequence
Particle swarm optimization
Tabu search

ABSTRACT

Hydrophobic–polar (HP) models are widely used to predict protein folding and hydrophobic interactions. Numerous optimization algorithms have been proposed to predict protein folding using the two-dimensional (2D) HP model. However, to obtain an optimal protein structure from the 2D HP model remains challenging. In this study, an algorithm integrating particle swarm optimization (PSO) and Tabu search (TS), named PSO–TS, was proposed to predict protein structures based on the 2D HP model. TS can help PSO to avoid getting trapped in a local optima and thus to remove the limitation of PSO in predicting protein folding by the 2D HP model. In this study, a total of 28 protein sequences were used to evaluate the accuracy of PSO–TS in protein folding prediction. The proposed PSO–TS method was compared with 15 other approaches for predicting short and long protein sequences. Experimental results demonstrated that PSO–TS provides a highly accurate, reproducible, and stabile prediction ability for the protein folding by the 2D HP model.

1. Introduction

In computational biology, protein structure prediction using amino acid sequences is currently receiving considerable attention because the folding of amino acid sequences determines the biological function of a protein [1]. Diseases such as amyloidosis, Parkinson disease, Alzheimer disease, and Creutzfeldt–Jakob disease are caused by abnormal protein folding processes [2]. Therefore, an accurate and effective method for protein folding prediction may contribute to the early detection and treatment of various diseases. Currently, protein structures are primarily detected through X-ray crystallography (XRC). However, XRC has many shortcomings, such as high cost, difficulty with protein crystallization, time-consuming analysis, and a highly technical skill requirement for pattern interpretation. Especially, some of the proteins have complex atomic structures and folding characteristics, which make their prediction by simulating the protein folding mechanism complicated and time consuming. Therefore, a more efficient analysis method to predict protein structures is required.

Computation by simulation is generally based on the protein primary structure, that is, the amino acid sequence. Among many of the protein structure prediction methods, the hydrophobic–polar (HP) protein folding model uses two- or three-dimensional (2D or 3D) lattices to simulate the protein folding on the basis of their amino acid sequences and thus provides information on the protein folding mechanism [3]. Using lattice models is considered a highly simplified method to simulate the protein folding process [3]. Protein central structures include hydrophobic (H) nonpolar amino acids, and surfaces comprise hydrophilic polar (P) amino acids. In general, the HP model calculates the strength of neighboring hydrophobic amino acids and assigns negative weights to these acids. A lower weight corresponds to a higher stability of the protein structure, thus representing a more homologous tertiary structure [3]. Although the HP model provides simple protein folding prediction, it has certain limitations. In some conditions, such as high-complexity protein applications, the HP model does not provide optimal solutions and results in a nondeterministic polynomial-time-hard (NP-hard) problem [4].

The hybrid optimization and random search algorithms have been widely used for protein folding problems [5]. Hoque et al. proposed a

* Corresponding author. Department of Chemical Engineering, I-Shou University, No.415, Jiangong Rd., Sanmin Dist., Kaohsiung City, 807, Taiwan.
** Corresponding author.
E-mail addresses: chyang@cc.kuas.edu.tw (C.-H. Yang), joe29681195@yahoo.com.tw (Y.-S. Lin), chuang@isu.edu.tw (L.-Y. Chuang), yudalinemail@gmail.com (Y.-D. Lin).

https://doi.org/10.1016/j.compbiomed.2019.103397
Received 15 May 2019; Received in revised form 19 August 2019; Accepted 19 August 2019
Available online 20 August 2019
0010-4825/ © 2019 Elsevier Ltd. All rights reserved.

hybrid genetic algorithm (HGA) for protein folding prediction using the 2D face-centered-cubic HP lattice model [6]. Böckenhauer et al. introduced a local-search neighborhood approach that employed so-called pull moves to determine optimal embedding of an amino acid sequence in a lattice [7]. Su et al. proposed a hybrid hill-climbing and genetic algorithm (HHGA) that used an elite-based reproduction strategy for protein folding prediction [8]. Smith compared several hybrid evolutionary learning optimization approaches, including the coevolving memetic algorithm (COMA), greedy COMA (GComa), steepest COMA (SComa), greedy COMA with randomly created rules (GRand), steepest COMA with random created rules (SRand), simple memetic algorithm (SMA), and genetic algorithm (GA), to protein structure prediction [9]. Yang et al. proposed a high-exploration particle swarm optimization (PSO)-based algorithm combined with local-search algorithms (HE–L–PSO) to predict protein folding using the 2D face-centered-cubic HP lattice model [10]. Chuang et al. combined a double-bottom chaotic map PSO method with a local-search approach (DBM–L–PSO) to predict protein folding structures [11]. In addition, Yang et al. introduced a hybrid algorithm that combines ion motion optimization with the greedy algorithm (IMOG) in the HP model for protein folding prediction [12]. In the HP model, the performance of a triangular lattice model is generally superior to that of a square lattice model [13]. Although several algorithms have provided nearly optimal solutions, the stability and accuracy of protein prediction remains room for improvement.

In this study, an algorithm integrating a local-search algorithm, Tabu search (TS) [14], and PSO algorithm [15] (named PSO–TS) was proposed to improve protein folding prediction. A total of 28 amino acid sequences with known structures were tested to evaluate the prediction accuracy of the proposed algorithm PSO–TS, in which a triangular lattice was used for the protein folding simulation.

2. Methods

2.1. Problem domain

Proteins and polypeptides, complex biological macromolecules, are composed of 20 basic amino acids. The biological function of proteins is determined by their native structure, which is based on their primary structure. Therefore, development of a highly efficient approach to predict protein folding is able to provide many advantageous in the study of protein biotechnology. Recently, many approaches have been proposed for protein structure prediction. The *ab initio* method provides direct prediction based on amino acid sequences instead of comparative modeling or fold recognition [3]. The HP model simulates the folding process of an amino acid sequence using lattice models, such as a triangular lattice, to predict protein folding. In this model, amino acids are classified as H or P type. However, extremely long-sequence folding and NP-hard problems require further investigation to obtain optimal solutions. In computational biology, this noteworthy limitation remains a challenge to be overcome [16].

According to Anfinsen's thermodynamic hypothesis, an amino acid sequence may fold into a particular tertiary structure with low free energy and resemble the real protein structure. This hypothesis was used to evaluate the HP protein folding model. When two hydrophobic amino acids are adjacent, a hydrophobic–hydrophobic (H–H) interaction occurs. As the number of H–H interactions increases, the stability and similarity of the predicted structure to natural protein folding increase.

2.2. Particle swarm optimization

In PSO, an available solution is represented as a particle. Assuming that N particles search for the solution in a D-dimensional feasible space, each particle has a current position $\vec{x}_i$ and velocity $\vec{v}_i$, where $i = 1, \ldots, N$. A fitness value indicates the quality of a particle. Fitness values can be evaluated using a candidate solution (i.e., particle position) from fitness function $fit(\vec{x}_i)$. The fitness value of the algorithm aims to lead particles towards the great search space. Each particle has the optimal position ($\vec{p}_i$) based on its searching experience, and the global optimal position ($\vec{p}_g$) is based on the searching experience of all particles. For PSO, the four primary operators of PSO are as follows: 1) initializing the population, 2) updating $\vec{p}_i$ and $\vec{p}_g$, 3) updating velocity and position of particles, and 4) evaluating termination criteria. The particle moves to a new position in the vector space with velocity $\vec{v}_i$, and the new position is calculated based on the basis of $\vec{p}_i$ and $\vec{p}_g$ vectors. Therefore, the particles can converge toward the optimal location in the search space.

2.3. Tabu search

TS, originally proposed by Glover [14], is a metaheuristic search algorithm that enables improvement of moves after encountering local optima. The TS algorithm can record short-term solutions to prevent duplicate search and thus avoid getting trapped in a local optimum and idle repetition of the search to find the previously visited solutions. Two components of TS are adapted to fit PSO algorithm.

2.4. PSO–TS algorithm

PSO provides excellent global search, and the TS algorithm records short-term results to prevent duplication of search pathways. In this study, the TS algorithm was integrated with PSO to exchange information between two particles to improve the protein folding prediction. If the fitness value does not change after information exchange, the optimal structure is maintained. However, in PSO, the population could easily become trapped in a local optimum solution. Thus, a Table list is used to store the current optimal solution for n iterations, and the information of particles are subsequently reset to prevent repetition of the same structure prediction. The flowchart of PSO–TS algorithm is shown in Fig. 1. First, the parameters are set and all particles are initialized. Then, the velocity and position of particles in PSO are continually updated. After updating the particles in each iteration, the TS algorithm is executed. If the result obtained from TS is not superior to that from PSO, then TS is executed. After a fixed number of iterations, $\vec{p}_g$ is stored in the Table list; subsequently, $\vec{p}_g$ information and some particles are initialized. The process of PSO–TS are detailed as below.

2.4.1. Step 1: Initialize particles

Six protein folding directions are encoded as {1, 2, …, 6} to indicate protein folding directions, and the particles are described as $\vec{x}_i = (x_{i1}, x_{i2}, \ldots, x_{id})$, where $d \in \{1, 2, \ldots, 6\}$, represents six neighbors in the 2D triangular lattice model. Because the direction of the first point toward with respect to the second point does not affect the protein folding, this movement direction is not recorded and is marked as "-". Fig. 2A displays the triangular lattice protein models. When an amino acid has a data value equals to 1, 2, 3, 4, 5 or 6, that represents the next subsequent amino acid folds to upper left, upper right, right, lower right, lower left, or left, respectively. For example, if the data value for the third amino acid is 1, the fourth amino acid folds to the upper left of the third amino acid. Fig. 2B and C displays the protein folding process and outcome.

For initializing the population, each position of particle $\vec{x}_i$ is randomly generated as a candidate solution within the search space, and velocity $\vec{v}_i$ is randomly generated within the parameters $[-V_{\max}, +V_{\min}]$. Each particle is randomly generated based on the basis of the aforementioned direction labels.

2.4.2. Step 2: Calculate fitness

The objective of protein folding problem is to determine a conformation with minimum energy. The values of $\vec{p}_i$ and $\vec{p}_g$ are obtained

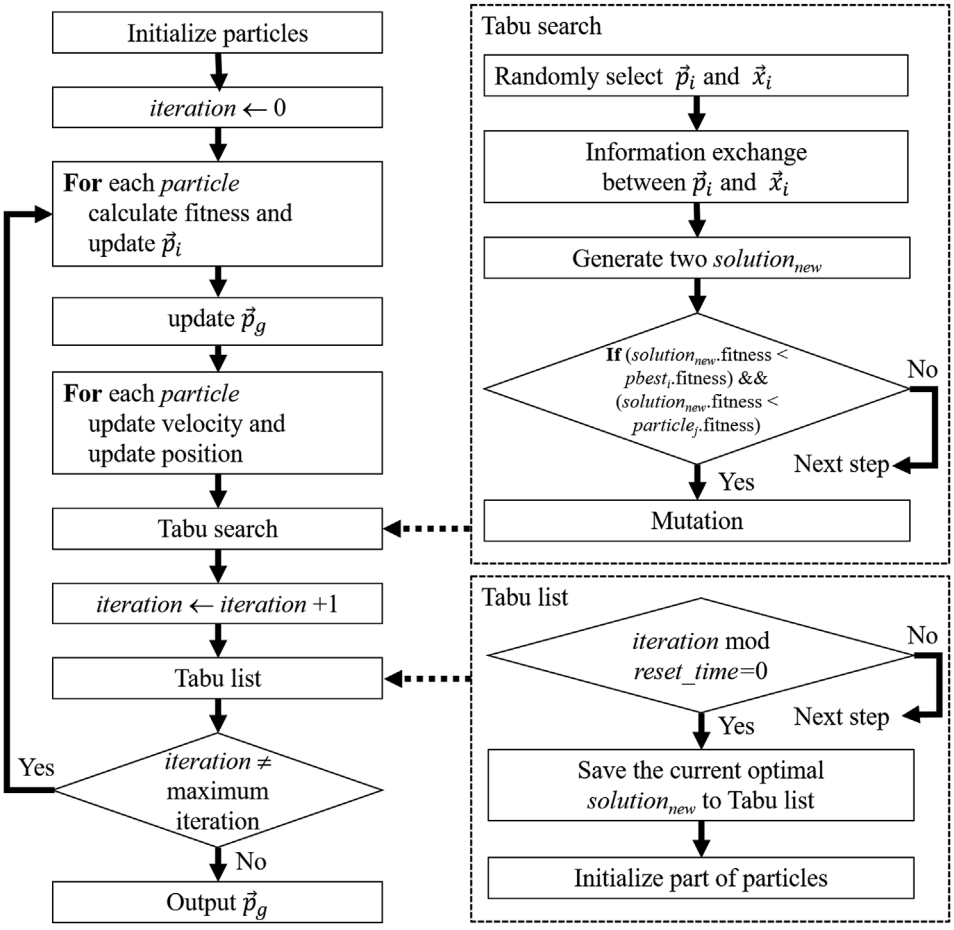

Fig. 1. Flowchart of PSO–TS algorithm.

after calculating all candidate solutions from the fitness function. Gibbs free energy is the most common parameter to estimate the stability of the predicted protein folding through H–H interactions. In the energy calculations of HP model, all amino acids are classified as hydrophobic nonpolar (H) or hydrophilic polar (P). Free energy is derived by counting the number of H–H interactions between pairs of nearest hydrophobic amino acids, and this parameter is used as the fitness value. The amino acid sequence $\boldsymbol{S} = s_1, s_2, \ldots, s_n$ is a protein chain with a length n. The objective of the problem is to determine the energy-minimizing conformation of $\boldsymbol{S}$, that is, to determine $c^* \in C(\boldsymbol{S})$ such that E

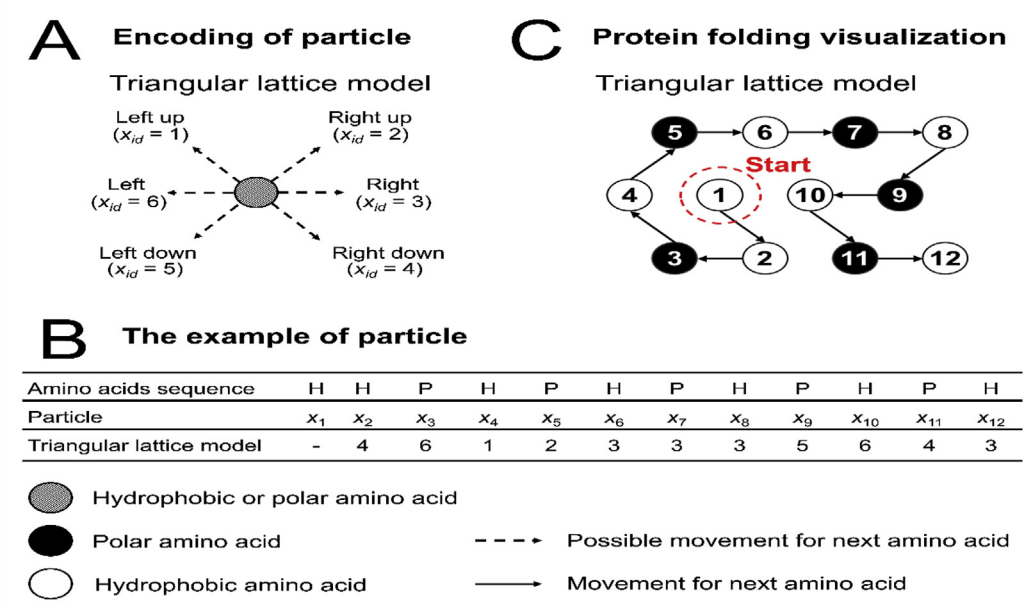

Fig. 2. Illustration of PSO encoding of protein folding direction for triangular lattice models.

$(c^*) = \min\{E(c) \mid c \in C\}$, where $C(\boldsymbol{S})$ is the set of all valid conformations for $\boldsymbol{S}$. The objective function can be mathematically defined as follows:

$$fitness = \sum_{i,j} \Delta\gamma_{ij} \cdot \sigma_{ij}, \tag{1}$$

where

$$\Delta\gamma_{ij} = \begin{cases} 1\ s_i \text{ and } s_j \text{ are adjacent but} \\ \quad \text{not connected amino acids} \\ 0\ \text{others} \end{cases} \text{ and} \tag{2}$$

$$\sigma_{ij} = \begin{cases} -1\ \text{the pair of } H \text{ and } H \text{ residues} \\ 0\ \text{others} \end{cases}. \tag{3}$$

2.4.3. Step 3: Update $\vec{p_i}$ and $\vec{p_g}$

Particles keep a record of their personal optimal position ($\vec{p_i}$) and the global optimal position ($\vec{p_g}$) when moving. If the fitness value of a particle $\vec{x_i}$ in the current iteration has a lower energy than the fitness value of $\vec{p_i}$, then the position and fitness value of $\vec{p_i}$ are replaced by the current position $\vec{x_i}$ and fitness value. When all updating processes of $\vec{p}$ is completed, each group (topology) selects the optimal solution with the lowest energy for $\vec{p_g}$. For each particle in the next iteration of PSO, $\vec{p_i}$ and $\vec{p_g}$ provide search directions towards energy minimization.

2.4.4. Step 4: Update velocity and position

Clerc and Kennedy proposed an effective particle updating function as follows:

$$\begin{cases} \vec{v_i} \leftarrow \chi\left(\begin{array}{l} \vec{v_i} + \vec{U}(0, \varphi_1) \otimes (\vec{p_i} - \vec{x_i}) \\ \quad + \vec{U}(0, \varphi_2) \otimes (\vec{p_g} - \vec{x_i}) \end{array}\right), \\ \vec{x_i} \leftarrow \vec{x_i} + \vec{v_i}, \end{cases} \tag{4}$$

where with $\varphi = \varphi_1 + \varphi_2 > 4$ and

$$\chi = \frac{2}{\varphi - 2 + \sqrt{\varphi^2 - 4\varphi}}, \tag{5}$$

where $\vec{U}(0, \varphi_i)$ is a vector of random numbers uniformly distributed in $[0, \varphi_i]$ for each iteration and for each particle. $\otimes$ is the component-wise multiplication. Using Clerc's constriction method, the parameter settings are defined as $\varphi = 4.1\varphi_1 = \varphi_2$, and the constant multiplier $\chi = 0.7298$. The position of particle $\vec{x}$ can be adjusted by adding the new velocity $\vec{v}$, such the adjusted positions of particles can provide the new protein foldings towards the objective of energy minimization.

2.4.5. Step 5: Tabu search

In Step 5.1 of the search process, information is exchanged between two particles to derive two predictions of the protein structure. If the updated prediction is superior to the original prediction, the original information is replaced with the new information. In this study, particles with optimal solutions from previous iterations are randomly selected to exchange information with random particles through a two-point crossover process. All resource information is exchanged between two randomly selected particles of different dimensions. For example, when particles of dimensions 3 and 6 are selected, all information from both of the particles is exchanged to obtain two new results. That is, the values of $\vec{p_i}$ and $\vec{x_i}$ for dimensions 3 and 6 are exchanged, yielding $solution_{new1}$ and $solution_{new2}$, respectively. In Step 5.2, the mutation probability is obtained for each dimension. The mutations within each dimension indicate the various folding possibilities. Each possible structure is evaluated to obtain the optimal structure. When a particular number of iterations is reached, the optimal structure solution is stored in the Tabu list, and the information of all particles is reset. Thus, subsequent iterations do not necessarily provide the same prediction solutions. The Table list records the short-term results to avoid duplication of the search process in the algorithm, thus obtaining optimal solutions beyond a local optimum solution. This feature can effectively compensate for the limitations of Step 5.2 in the search process, because the optimal solutions are recorded after few iterations, and the information of all particles is reset to prevent the particles from moving in previously searched directions.

2.4.6. Step 6: Evaluate termination criteria

The process is stopped after the maximum number of iterations is reached.

2.5. Parameter setting

The numbers of particles and iterations for the PSO algorithm are set to 100 and 200, respectively [17]. Because a triangular lattice model is used to predict folding in six directions, the updated speeds $V_{\min}$ and $V_{\max}$ are -5 and 5, respectively. In addition, the minimum and maximum search ranges are 1 and 6, respectively. The Tabu list stores the optimal solution of a set of 10 iterations, after which the information is reset.

3. Results

3.1. Datasets

The prediction accuracy and reproducibility of the proposed algorithm using triangular lattice models were tested. For this purpose, benchmark HP sequences and their optimum values were obtained from the previous researches [8,18] (Table 1).

3.2. Evaluation of PSO–TS using three topologies

The accuracy of protein structure prediction using PSO–TS was evaluated by three topologies: ring, star, and completely connected topologies. A protein structure with a smaller fitness value indicates having a greater free energy and, therefore, is a more accurate prediction. Topology can affect the search strategy of population in each iteration. Among the ring, star, and completely connected topologies, PSO–TS performed the best accuracy for the completely connected topology (data not shown). Therefore, the completely connected topology is the optimal topology for protein structure prediction using PSO–TS and the 2D HP model.

3.3. Comparison of prediction accuracies of PSO–TS and other algorithms

As shown in Fig. 3, the prediction accuracies of PSO-TS were compared with SGA [6], HGA [6], TS [7], ERS–GA, HHGA [8] and IMOG [12] for triangular lattice models. All algorithms, excluding SGA and HGA, exhibited high performance in searching for sequences 1–3. PSO–TS outperformed the other algorithms for sequences 5–8. Fig. 3B showed the comparison of the average optimal results obtained from 25 experiments. For all sequences, the performance of PSO–TS was superior to that of the other algorithms. The Wilcoxon signed-rank test was conducted to compare the performance of PSO–TS and other algorithms for eight sequences (Fig. 3C and D). A p value of less than 0.05 indicates significant superiority of PSO–TS to other methods. R^- represents the degree of inferiority of PSO–TS with respect to other methods. PSO–TS was not inferior to the other methods for any test sequence.

Furthermore, the reproducibility of PSO–TS prediction was compared with that of 11 methods—namely, GComa [9], SComa [9], GRand [9], SRand [9], SMA [9], GA [9], high-exploration PSO (HEPSO) [10], HLS [10], DBM–L–PSO [11], HE–L–PSO [10], and IMOG [12]—for 25 iterations of each sequence (Fig. 4). The reproducibility of PSO–TS was superior to that of the 11 methods. A significant superiority ($p < 0.01$) was observed between PSO–TS and the other methods

Table 1
HP sequences of the triangular lattice model.

Seq. code	Len	Amino acid sequences	E*(Tra)
1–1	20	$(HP)^2PH(HP)^2(PH)^2HP(PH)^2$	−15
1–2	24	$H^2P^2(HP^2)^6H^2$	−17
1–3	25	$P^2HP^2(H^2P^4)^3H^2$	−12
1–4	36	$P(P^2H^2)^2P^5H^5(H^2P^2)^2P^2H(HP^2)^2$	−24
1–5	48	$P^2H(PH^3)^2P^5H^{10}P^6(H^2P^2)^2HP^2H^5$	−43
1–6	50	$H^2(PH)^3PH^4PH(P^3H)^2P^4(HP^3)^2HPH^4(PH)^3PH^2$	−41
1–7	60	$P(PH^3)^2H^5P^3H^{10}PHP^3H^{12}P^4H^6PH^2PHP$	N.D.
1–8	64	$H^{12}(PH)^2((P^2H^2)^2P^2H)^3(PH)^2H^{11}$	N.D.
2–1	12	$H(HP)^2H$	−11
2–2	14	$H^2P^2(HP)^5$	−11
2–3	14	$H^2P^2HP(PH)^4$	−11
2–4	16	$H^2PH(P^2H)^4$	−11
2–5	16	$H(HP^2)^2(HP)^3PHP$	−11
2–6	17	$H(HP^2)^5H$	−11
2–7	17	$H(HP)^7H^2$	−17
2–8	20	$H(HP^2)^2(HP)^3(PH)^3H$	−17
2–9	20	$H(HP)^5(PHP)^2PH^2$	−17
2–10	21	$H^2P^2(HP^2HP)^3H^2$	−17
2–11	21	$H^2P(HP^2)^2(HP)^3PHP^2H^2$	−17
2–12	21	$H^2P^2(HP)^3(PH)^2(P^2H)^2H$	−17
2–13	22	$H(HP^2)^2(HP)^3PH(P^2H)^2H$	−17
2–14	23	$H^2(HP)^9H^3$	−25
2–15	24	$H(HP^2)^7H^2$	−17
2–16	24	$H^2(HP)^3(PH)^7H^2$	−25
2–17	24	$H^2(HP)^4(PH)^6H^2$	−25
2–18	30	$H^3(P^2H)^4(PHP)^2(HP^2)^2H^3$	−25
2–19	30	$H^3(P^2H)^3(PHP)^2(HP^2)^3H^3$	−25
2–20	37	$H^3(P^2H)^3(PH)^2(P^2H)^3P^4(PH)^2PH^3$	−29

Seq. code: sequence code. Len: number of residues of the sequence. *H*: hydrophobic. *P*: polar. H^n: *n Hs*. P^n: *n Ps*. HP^n: *n HPs*. PH^n: *n PHs*. Sqr: square lattice model. Tra: triangular lattice model. E*: negative value indicates the number of hydrophobic interactions. Lower values indicate greater predicted protein folding stability based on more hydrophobic interactions. The best results for triangular lattice models 1-1 to 1–8 were reported in Ref. [18]. The best results for triangular lattice models 2-1 to 2–20 have been reported in Refs. [8,13]. "N.D." indicates no determination.

(Fig. 4B), indicating that PSO–TS improved the protein folding prediction. Fig. 5 displays the optimal values obtained using the seven algorithms for sequences 14–20 and 25 iterations. PSO–TS exhibited the best performance among the seven algorithms.

4. Discussion

In this study, we proposed PSO–TS algorithm which combines two strategies for local search to obtain a superior protein structure in 25 experiments. The prediction abilities of PSO–TS were compared with that of other available algorithms in terms of the protein folding prediction with a lower free energy. Moreover, the reproducibility of PSO–TS was compared with that of other available algorithms by calculating the predicted optimal protein folding structures in 25 experiments. The results indicated that PSO–TS algorithm was more effective in identifying superior protein structures compared with the other 11 algorithms for 28 amino acid sequences with known structures.

PSO has demonstrated favorable convergence speed, global optimality, and accuracy in many applications. Although PSO has been widely used to predict protein folding, its local-search ability remains unsatisfactory, particularly for the prediction of complex protein structures. Protein folding prediction requires an effective method for searching local solutions. Some local-search methods, such as the hill-climbing algorithm, provide high-performance on protein folding prediction; however, many research results have indicated that the prediction abilities of local-search methods remain insufficient [6]. In this study, the proposed algorithm, PSO–TS, combined the features of PSO and TS for local search and demonstrated superior protein structure prediction compared with other algorithms in 25 experiments. In the prediction of 28 amino acid sequences with known structures, PSO–TS algorithm was more effective than 11 other algorithms.

The HP model predicts protein folding by determining the optimal structure based on the number of neighboring hydrophobic amino acids under the assumption that the hydrophilic reaction contributes to the free energy of folded proteins. Numerous algorithms have been applied on local search to improve the accuracy of protein folding prediction, including SGA [6], HGA [6], ERS–GA [8], HHGA [8], GComa [9], SComa [9], GRand [9], SRand [9], SMA [9], HEPSO [10], DBM–L–PSO [11], and HE–L–PSO [10]. However, the local-search strategies of these algorithms may lead their populations to become trapped in local optima. In PSO–TS, PSO predicts protein folding, and Steps 1 and 2 of TS as well as the repetitions of Table list operations improve the PSO search. Because particles may exhibit poor adaptability in early iterations, Step 1 of the search process may not improve the accuracy of protein folding prediction. When the protein structure is not dense, Step 1 of the search process is crucial to improve the prediction accuracy of protein folding structure. In the search process, if Step 1 does not improve the adaptability of particles, Step 2 is performed to improve the prediction of protein folding structure. Therefore, the number of neighboring hydrophobic amino acids in the predicted protein structure can be increased by increasing the iteration number. Step 1 of the search process can effectively detect a long-sequence protein structure using an operation involving the exchange of substructures between two dense protein folding structures. In the search space, particles can be aggregated into a local region with numerous generations. PSO may encounter the local optimum problem, which hinders improvement of prediction, particularly when numerous adjacent hydrophobic amino

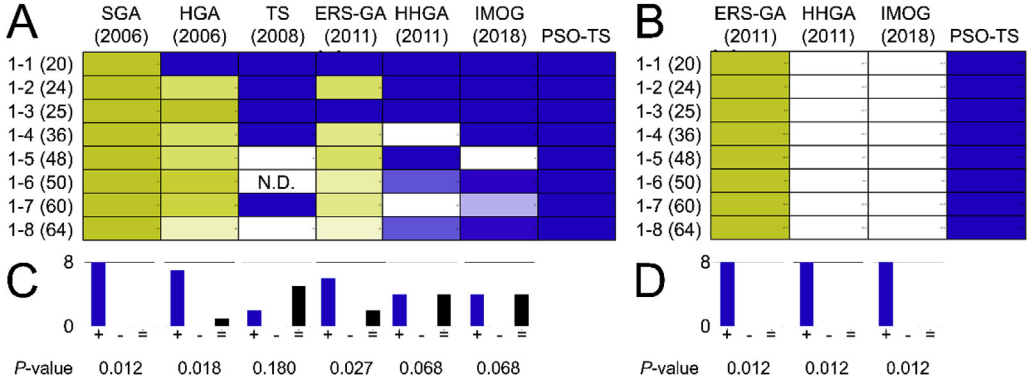

Fig. 3. (A) Comparison of performance of triangular lattice models for HP sequences listed in Table 1 (B) Comparison of the prediction stability of PSO–TS and other algorithms. (C) Comparison of performance of PSO–TS and other algorithms among eight datasets using Wilcoxon signed-rank test. (D) Comparison of prediction stability of PSO–TS and other algorithms among eight datasets using Wilcoxon signed-rank test. The *Y*-axis represents the number of datasets; "−," the degree to which PSO–TS is inferior to other algorithms; "+," the degree to which PSO–TS is superior to other algorithms; " = ," the degree to which PSO–TS is equal to other algorithms; and "N.D.," "not determined."

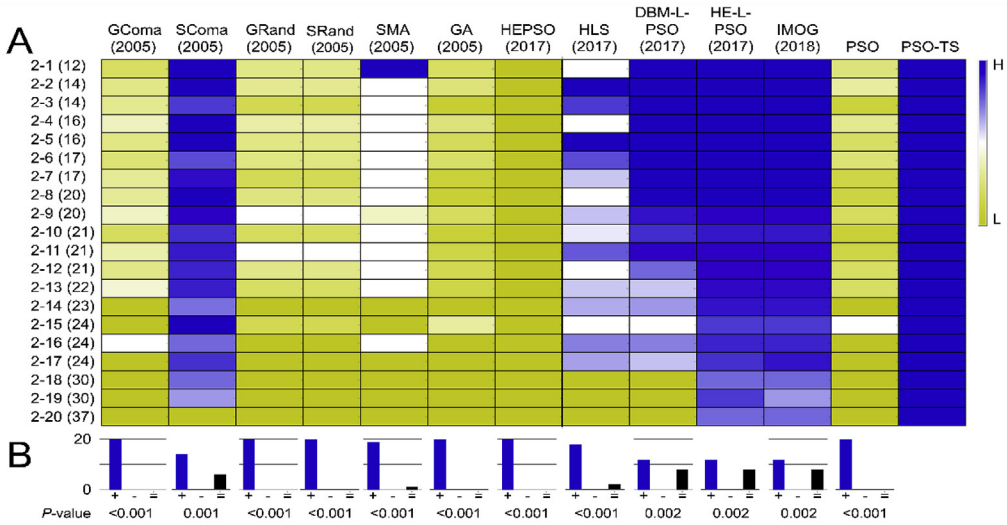

Fig. 4. Comparison of reproducibility of optimal prediction results among 25 experiments. (A) Reproducibility of optimal prediction results among 25 experiments. Darker blue (H) denotes superior implementation, and darker green (L) denotes weak implementation in the corresponding region. (B) Comparison of reproducibility of optimal prediction results of PSO–LS and other algorithms for 20 datasets by Wilcoxon signed-rank test. The *Y*-axis is the number of datasets. "−," the degree to which PSO–TS is inferior to other algorithms; "+," the degree to which PSO–TS is superior to other algorithms; " = ," the degree to which PSO–TS is equal to other algorithms. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

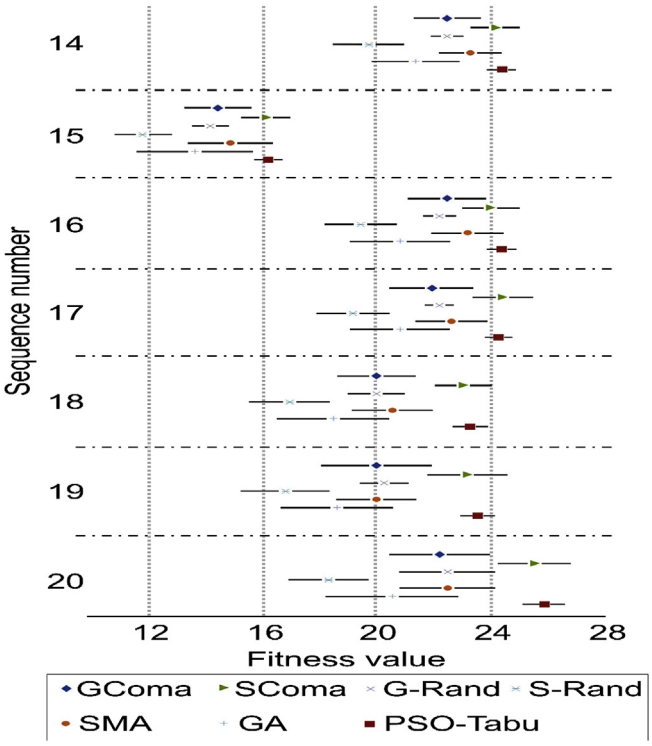

Fig. 5. Optimal prediction results obtained using seven algorithms for sequences 14–20 after 25 iterations.

acids are present. TS ensures that solutions are escape local optima by enabling long-distance movement in the search space. Moreover, the Tabu list is used to stop the particles from moving along previous search paths by updating $\vec{p}_i$ after each set of a specified number of iterations. Thus, the TS steps provide improved accuracy of prediction solutions for new local particles.

The advantages of the proposed algorithm, PSO–TS, in this study are summarized as follows:

1) High reproducibility: PSO–TS provided the highest reproducibility of prediction score among all algorithms for 25 iterations of the same sequence.
2) High performance in the prediction of long-sequence protein folding: Structure prediction by PSO–TS for long amino acid sequences was superior to that by other algorithms.
3) High stability of protein folding prediction: In general, the success of protein folding prediction for long amino acid sequences is lower than that for short sequences. As the length of the sequence increases, the complexity of the protein folding space around local minima increases. The prediction ability of PSO–TS was superior to that of other methods. These results indicate that PSO–TS provides an improved triangular lattice model for protein folding prediction.

In this study, PSO-TS continues to possess a number of limitations. The length of the amino acid sequence can affect the performance of PSO–TS. An amino acid sequence of length n contains 6^n protein folding combinations in a 2D HP model with six possible folding directions. When n increases, the number of available protein folding combinations exponentially increases. Thus, a large search space can increase the difficulty of search in PSO–TS for determining the optimal protein folding. We suggested that the number of particles and iterations for the PSO–TS algorithm can be set to large numbers when the length of the amino acid sequence is large. Furthermore, the fitness value of PSO–TS is determined on the basis of relative free energy; however, for precise comparison, a fitness function based on absolute free energy for predicting protein folding structure can benefit future investigation.

5. Conclusions

This study presents an algorithm that integrates TS with PSO to predict protein structures in the HP model. PSO–TS demonstrated outstanding performance in predicting protein folding, particularly in terms of the stability of the evolutionary algorithm.

Authors' contributions

C-HY conceived the study and participated in the design of the algorithm. Y-SL participated in the design of the algorithm and writing of the program. L-YC and Y-DL designed the study, performed the analyses, and drafted the manuscript.

Conflicts of interest

The authors declare that they have no competing interests.

Acknowledgments

This work was partly supported by the Ministry of Science and

Technology, Taiwan, R.O.C. (108-2811-E-992-502-, 108-2221-E-992-031-MY3, 108-2221-E-214 -019-MY3, and 108-2221-E-992-031-MY3).

List of Abbreviations

HP	hydrophobic–polar
2D	two-dimensional
TS	Tabu search
NP-hard	nondeterministic polynomial-time-hard
HGA	hybrid genetic algorithm
HHGA	hybrid of hill-climbing and genetic algorithm
COMA	coevolving memetic algorithm
GComa	greedy COMA
SComa	steepest COMA
GRand	greedy COMA with randomly created rules
SRand	steepest COMA with randomly created rules
SMA	greedy-ascent memetic algorithm
GA	genetic algorithm
PSO	particle swarm optimization
HE–L–PSO	high-exploration particle swarm optimization-based algorithm combined with a local-search algorithm
DBM–PSO	double-bottom chaotic map particle swarm optimization
IMOG	ion motion optimization with a greedy algorithm
PSO–TS	particle swarm optimization–Tabu search
H–H	hydrophobic–hydrophobic
HEPSO	high-exploration PSO

References

[1] C. Anfinsen, The formation and stabilization of protein structure, Biochem. J. 128 (4) (1972) 737.
[2] L.C. Walker, H. LeVine, The cerebral proteopathies: neurodegenerative disorders of protein conformation and assembly, Mol. Neurobiol. 21 (1–2) (2000) 83–95.
[3] K.A. Dill, Theory for the folding and stability of globular proteins, Biochemistry 24 (6) (1985) 1501–1509.
[4] A. Madain, A.L.A. Dalhoum, A. Sleit, Application of local rules and cellular automata in representing protein translation and enhancing protein folding approximation, Progress in Artificial Intelligence 7 (3) (2018) 225–235.
[5] N.D. Jana, J. Sil, Protein structure prediction in 2D HP lattice model using differential evolutionary algorithm, Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 Held in Visakhapatnam, India, January 2012, Springer, 2012, pp. 281–290.
[6] M.T. Hoque, M. Chetty, L.S. Dooley, A Hybrid Genetic Algorithm for 2D FCC Hydrophobic-Hydrophilic Lattice Model to Predict Protein Folding, Advances in Artificial Intelligence, Springer, 2006, pp. 867–876.
[7] H.-J. Böckenhauer, A.Z.M.D. Ullah, L. Kapsokalivas, K. Steinhöfel, A Local Move Set for Protein Folding in Triangular Lattice Models, Algorithms in Bioinformatics, Springer, 2008, pp. 369–381.
[8] S.C. Su, C.J. Lin, C.K. Ting, An effective hybrid of hill climbing and genetic algorithm for 2D triangular protein structure prediction, Proteome Sci. 9 (Suppl 1) (2011) S19.
[9] J. Smith, The Co-evolution of Memetic Algorithms for Protein Structure Prediction, Recent Advances in Memetic Algorithms, Springer, 2005, pp. 105–128.
[10] C.H. Yang, Y.S. Lin, L.Y. Chuang, H.W. Chang, A particle swarm optimization-based approach with local search for predicting protein folding, J. Comput. Biol. 24 (10) (2017) 981–994.
[11] L.Y. Chuang, Y.D. Lin, C.H. Yang, High-performance computing for protein fold prediction, 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, IEEE, 2017, pp. 1–6.
[12] C.H. Yang, K.C. Wu, Y.S. Lin, L.Y. Chuang, H.W. Chang, Protein folding prediction in the HP model using ions motion optimization with a greedy algorithm, BioData Min. 11 (1) (2018) 17.
[13] S.P. Dubey, S. Balaji, N.G. Kini, M.S. Kumar, A comparative study of various metaheuristic algorithms for Ab initio protein structure prediction on 2D hydrophobic-polar model, Proceedings of Fifth International Conference on Soft Computing for Problem Solving, Springer, 2016, pp. 387–399.
[14] F. Glover, Tabu search-part I, ORSA J. Comput. 1 (3) (1989) 190–206.
[15] R. Mendes, J. Kennedy, J. Neves, The fully informed particle swarm: simpler, maybe better, IEEE Trans. Evol. Comput. 8 (3) (2004) 204–210.
[16] A. Bechini, On the characterization and software implementation of general protein lattice models, PLoS One 8 (3) (2013) e59504.
[17] J. Kennedy, Particle Swarm Optimization, Encyclopedia of Machine Learning, Springer, 2011, pp. 760–766.
[18] M.K. Islam, M. Chetty, M. Murshed, Novel local improvement techniques in clustered memetic algorithm for protein structure prediction, IEEE Congress on Evolutionary Computation, 2011, pp. 1003–1011.